Variational Bayesian Approximation methods for inverse problems

Ali Mohammad-Djafari
Laboratoire des Signaux et Systèmes,
UMR8506 CNRS-SUPELEC-UNIV PARIS SUD 11
SUPELEC, 91192 Gif-sur-Yvette, France
http://lss.supelec.free.fr

Email: djafari@lss.supelec.fr
http://djafari.free.fr

Content

1. General inverse problem

$$g(t) = \mathcal{H} f(t) + \epsilon(t), \quad t \in [1, \cdots, T]$$

$$g(\boldsymbol{r}) = \mathcal{H} f(\boldsymbol{r}) + \epsilon(\boldsymbol{r}), \quad \boldsymbol{r} = (x, y) \in R^2$$

- f unknown quantity (input)
- $\mathcal{H}$ Forward operator:
  (Convolution, Radon, Fourier or any Linear operator)
- g observed quantity (output)
- ϵ represents the errors of modeling and measurement

Discretization:

$$\boldsymbol{g} = \boldsymbol{H} \boldsymbol{f} + \boldsymbol{\epsilon}$$

- Forward operation $\boldsymbol{H} \boldsymbol{f}$
- Adjoint operation $\boldsymbol{H}' \boldsymbol{g}$: $\quad < \boldsymbol{H}' \boldsymbol{g}, \boldsymbol{f} > = < \boldsymbol{H} \boldsymbol{f}, \boldsymbol{g} >$
- Inverse operation (if exists) $\boldsymbol{H}^{-1} \boldsymbol{g}$

2. General Bayesian Inference

- Bayesian inference:

$$p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta}) = \frac{p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)}{p(\boldsymbol{g}|\boldsymbol{\theta})}$$

  with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$

- Point estimators:
  Maximum A Posteriori (MAP) or Posterior Mean (PM) $\longrightarrow \widehat{\boldsymbol{f}}$

- Full Bayesian inference:
- Simple prior models: $p(\boldsymbol{f}|\boldsymbol{\theta}_2)$

$$q(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta})$$

- Prior models with hidden variables: $p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta}_2)\; p(\boldsymbol{z}|\boldsymbol{\theta}_3)$

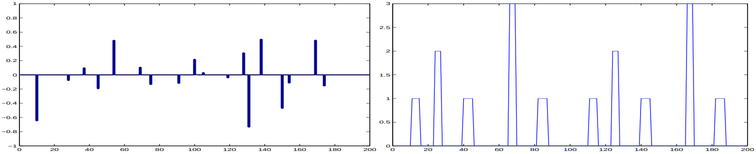
$$q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta}_2)\, p(\boldsymbol{z}|\boldsymbol{\theta}_3)\, p(\boldsymbol{\theta})$$

Two main steps in the Bayesian approach

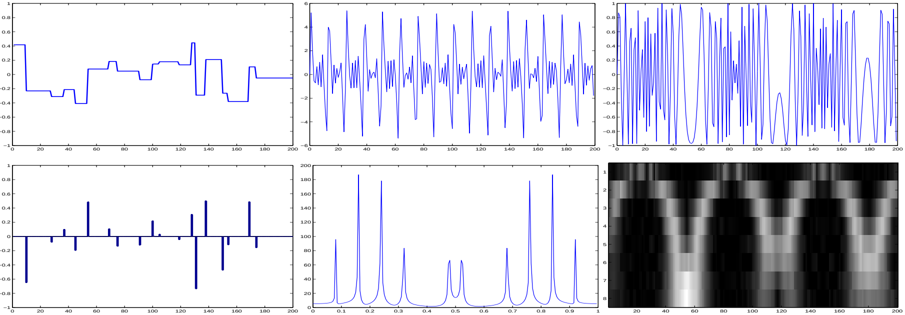
- Prior modeling
  - Separable:
    Gaussian, Generalized Gaussian, Gamma,
    mixture of Gaussians, mixture of Gammas, ...
  - Markovian: Gauss-Markov, GGM, ...
  - Separable or Markovian with hidden variables
    (contours, region labels)
- Choice of the estimator and computational aspects
  - MAP, Posterior mean, Marginal MAP
  - MAP needs optimization algorithms
  - Posterior mean needs integration methods
  - Marginal MAP and Hyperparameter estimation need
    integration and optimization
  - Approximations:
    - Gaussian approximation (Laplace)
    - Numerical exploration MCMC
    - Variational Bayes (Separable approximation)

3. Sparsity enforcing prior models

- Sparse signals: Direct sparsity



- Sparse signals: Sparsity in a Transform domaine



3. Sparsity enforcing prior models

- Simple heavy tailed models:
  - Generalized Gaussian, Double Exponential
  - Student-t, Cauchy
  - Elastic net

  - Symmetric Weibull, Symmetric Rayleigh
  - Generalized hyperbolic

- Hierarchical mixture models:
  - Mixture of Gaussians
  - Bernoulli-Gaussian

  - Mixture of Gammas
  - Bernoulli-Gamma
  - Mixture of Dirichlet
  - Bernoulli-Multinomial

Simple heavy tailed models

- Generalized Gaussian, Double Exponential

$$p(\boldsymbol{f}|\gamma,\beta)=\prod_j \mathcal{GG}(f_j|\gamma,\beta) \propto \exp\left\{-\gamma\sum_j |f_j|^\beta\right\}$$

$\beta=1$ Double exponential or Laplace.
$0<\beta\leq 1$ are of great interest for sparsity enforcing.

- Student-t and Cauchy models

$$p(\boldsymbol{f}|\nu)=\prod_j \mathcal{S}t(f_j|\nu) \propto \exp\left\{-\frac{\nu+1}{2}\sum_j \log\left(1+f_j^2/\nu\right)\right\}$$

Cauchy model is obtained when $\nu=1$.

- Elastic net prior model

$$p(\boldsymbol{f}|\nu)=\prod_j \mathcal{EN}(f_j|\nu) \propto \exp\left\{-\sum_j (\gamma_1|f_j|+\gamma_2 f_j^2)\right\}$$

Mixture models

- Mixture of two Gaussians (MoG2) model

$$p(\boldsymbol{f}|\lambda, v_1, v_0) = \prod_j \left(\lambda \mathcal{N}(f_j|0, v_1) + (1-\lambda)\mathcal{N}(f_j|0, v_0)\right)$$

- Bernoulli-Gaussian (BG) model

$$p(\boldsymbol{f}|\lambda, v) = \prod_j p(f_j) = \prod_j \left(\lambda \mathcal{N}(f_j|0, v) + (1-\lambda)\delta(f_j)\right)$$

- Mixture of Gammas

$$p(\boldsymbol{f}|\lambda, v_1, v_0) = \prod_j \left(\lambda \mathcal{G}(f_j|\alpha_1, \beta_1) + (1-\lambda)\mathcal{G}(f_j|\alpha_2, \beta_2)\right)$$

- Bernoulli-Gamma model

$$p(\boldsymbol{f}|\lambda, \alpha, \beta) = \prod_j \left[\lambda \mathcal{G}(f_j|\alpha, \beta) + (1-\lambda)\delta(f_j)\right]$$

MAP, Joint MAP

- Inverse problems: $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}$
- Posterior law:

$$p(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)$$

- Examples:
  Gaussian noise, Gaussian prior and MAP:

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f})\} \text{ with } J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{f}\|_2^2$$

  Gaussian noise, Double Exponential prior and MAP:

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f})\} \text{ with } J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{f}\|_1$$

- Full Bayesian: Joint Posterior:

$$p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta})$$

- Joint MAP:

$$(\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\theta}}) = \arg\max_{(\boldsymbol{f}, \boldsymbol{\theta})} \{p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})\}$$

Marginal MAP and PM estimates

- Marginal MAP: $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{p(\boldsymbol{\theta}|\boldsymbol{g})\}$ where

$$p(\boldsymbol{\theta}|\boldsymbol{g}) = \int p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) \, \mathrm{d}\boldsymbol{f} = \int p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1) \, p(\boldsymbol{f}|\boldsymbol{\theta}_2) \, \mathrm{d}\boldsymbol{f}$$

  and then $\widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \left\{ p(\boldsymbol{f}|\widehat{\boldsymbol{\theta}}, \boldsymbol{g}) \right\}$

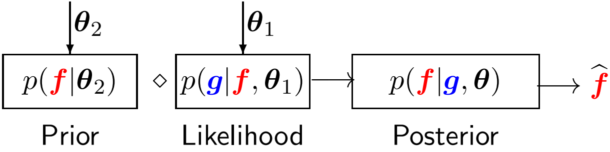
- Posterior Mean: $\widehat{\boldsymbol{f}} = \int \boldsymbol{f} \, p(\boldsymbol{f}|\widehat{\boldsymbol{\theta}}, \boldsymbol{g}) \, \mathrm{d}\boldsymbol{f}$

- EM and GEM Algorithms

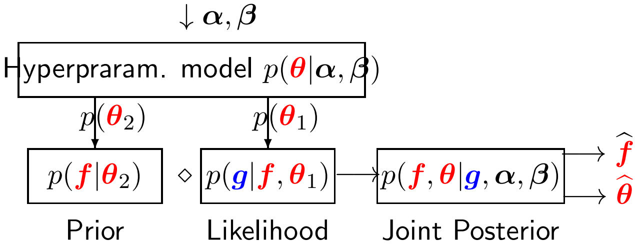
- Variational Bayesian Approximation:
  Approximate $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})$ by $q(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f}|\boldsymbol{g}) \, q_2(\boldsymbol{\theta}|\boldsymbol{g})$
  and then continue computations.

Summary of Bayesian estimation 1

- Simple Bayesian Model and Estimation



- Full Bayesian Model and Hyperparameter Estimation



Summary of Bayesian estimation 2

- Marginalization for Hyperparameter Estimation

$$\boxed{p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{p(\boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \widehat{\boldsymbol{\theta}} \longrightarrow \boxed{p(\boldsymbol{f}|\widehat{\boldsymbol{\theta}},\boldsymbol{g})} \longrightarrow \widehat{\boldsymbol{f}}$$

Joint Posterior   Marginalize over $\boldsymbol{f}$

- Variational Bayesian Approximation

$$\boxed{p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})} \longrightarrow \boxed{\text{Variational Bayesian Approximation}} \begin{array}{l} \longrightarrow q_1(\boldsymbol{f}) \longrightarrow \widehat{\boldsymbol{f}} \\ \longrightarrow q_2(\widehat{\boldsymbol{\theta}}) \longrightarrow \widehat{\boldsymbol{\theta}} \end{array}$$

Variational Bayesian Approximation

- Full Bayesian: $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta})$
- Approximate $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})$ by $q(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f}|\boldsymbol{g})\, q_2(\boldsymbol{\theta}|\boldsymbol{g})$ and then continue computations.
- Criterion $\mathrm{KL}(q(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) : p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}))$
- $\mathrm{KL}(q : p) = \int\int q \ln q/p = \int\int q_1 q_2 \ln \frac{q_1 q_2}{p}$
- Iterative algorithm $q_1 \longrightarrow q_2 \longrightarrow q_1 \longrightarrow q_2, \cdots$

$$\left\{ \begin{array}{ll} \widehat{q}_1(\boldsymbol{f}) & \propto \exp\left\{ \langle \ln p(\boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}; \mathcal{M}) \rangle_{\widehat{q}_2(\boldsymbol{\theta})} \right\} \\ \widehat{q}_2(\boldsymbol{\theta}) & \propto \exp\left\{ \langle \ln p(\boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}; \mathcal{M}) \rangle_{\widehat{q}_1(\boldsymbol{f})} \right\} \end{array} \right.$$

$p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})$ $\longrightarrow$ Variational Bayesian Approximation

$\longrightarrow q_1(\boldsymbol{f}) \longrightarrow \widehat{\boldsymbol{f}}$

$\longrightarrow q_2(\widehat{\boldsymbol{\theta}}) \longrightarrow \widehat{\boldsymbol{\theta}}$

BVA: Choice of the critrion

$$\mathsf{KL}(q:p) = \iint q(\boldsymbol{f},\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g};\mathcal{M})}{q(\boldsymbol{f},\boldsymbol{\theta})} \, \mathsf{d}\boldsymbol{f} \, \mathsf{d}\boldsymbol{\theta}$$

$$\begin{aligned} p(\boldsymbol{g}|\mathcal{M}) &= \iint q(\boldsymbol{f},\boldsymbol{\theta}) \frac{p(\boldsymbol{g},\boldsymbol{f},\boldsymbol{\theta}|\mathcal{M})}{q(\boldsymbol{f},\boldsymbol{\theta})} \, \mathsf{d}\boldsymbol{f} \, \mathsf{d}\boldsymbol{\theta} \\ &\geq \iint q(\boldsymbol{f},\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{g},\boldsymbol{f},\boldsymbol{\theta}|\mathcal{M})}{q(\boldsymbol{f},\boldsymbol{\theta})} \, \mathsf{d}\boldsymbol{f} \, \mathsf{d}\boldsymbol{\theta}. \end{aligned}$$

$$\mathcal{F}(q) = \iint q(\boldsymbol{f},\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{g},\boldsymbol{f},\boldsymbol{\theta}|\mathcal{M})}{q(\boldsymbol{f},\boldsymbol{\theta})} \, \mathsf{d}\boldsymbol{f} \, \mathsf{d}\boldsymbol{\theta}$$

$$p(\boldsymbol{g}|\mathcal{M}) = \mathcal{F}(q) + \mathsf{KL}(q:p)$$

BVA: Separable Approximation

$$p(\boldsymbol{g}|\mathcal{M}) = \mathcal{F}(q) + \mathsf{KL}(q:p)$$

$$q(\boldsymbol{f},\boldsymbol{\theta}) = q_1(\boldsymbol{f})\; q_2(\boldsymbol{\theta})$$

$$(\widehat{q}_1,\widehat{q}_2) = \arg\min_{(q_1,q_2)} \{\mathsf{KL}(q_1 q_2 : p)\} = \arg\max_{(q_1,q_2)} \{\mathcal{F}(q_1 q_2)\}$$

$\mathsf{KL}(q_1 q_2 : p)$ is convexe wrt q_1 when q_2 is fixed and vise versa:

$$\left\{\begin{array}{ll} \widehat{q}_1 & = \arg\min_{q_1} \{\mathsf{KL}(q_1 \widehat{q}_2 : p)\} = \arg\max_{q_1} \{\mathcal{F}(q_1 \widehat{q}_2)\} \\ \widehat{q}_2 & = \arg\min_{q_2} \{\mathsf{KL}(\widehat{q}_1 q_2 : p)\} = \arg\max_{q_2} \{\mathcal{F}(\widehat{q}_1 q_2)\} \end{array}\right.$$

$$\left\{\begin{array}{ll} \widehat{q}_1(\boldsymbol{f}) & \propto \exp\left\{\langle \ln p(\boldsymbol{g},\boldsymbol{f},\boldsymbol{\theta};\mathcal{M})\rangle_{\widehat{q}_2(\boldsymbol{\theta})}\right\} \\ \widehat{q}_2(\boldsymbol{\theta}) & \propto \exp\left\{\langle \ln p(\boldsymbol{g},\boldsymbol{f},\boldsymbol{\theta};\mathcal{M})\rangle_{\widehat{q}_1(\boldsymbol{f})}\right\} \end{array}\right.$$

BVA: Choice of family of laws q_1 and q_2

- Case 1 : $\longrightarrow$ Joint MAP

$$\begin{cases} \widehat{q}_1(\boldsymbol{f}|\widetilde{\boldsymbol{f}}) & = \delta(\boldsymbol{f} - \widetilde{\boldsymbol{f}}) \\ \widehat{q}_2(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}) & = \delta(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) \end{cases} \longrightarrow \begin{cases} \widetilde{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \left\{ p(\boldsymbol{f}, \widetilde{\boldsymbol{\theta}}|\boldsymbol{g}; \mathcal{M}) \right\} \\ \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\widetilde{\boldsymbol{f}}, \boldsymbol{\theta}|\boldsymbol{g}; \mathcal{M}) \right\} \end{cases}$$

- Case 2 : $\longrightarrow$ EM

$$\begin{cases} \widehat{q}_1(\boldsymbol{f}) & \propto p(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{g}) \\ \widehat{q}_2(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}) & = \delta(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) \end{cases} \longrightarrow \begin{cases} Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \langle \ln p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}; \mathcal{M}) \rangle_{q_1(\boldsymbol{f}|\widetilde{\boldsymbol{\theta}})} \\ \widetilde{\boldsymbol{\theta}} \qquad = \arg\max_{\boldsymbol{\theta}} \left\{ Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) \right\} \end{cases}$$

- Appropriate choice for inverse problems

$$\begin{cases} \widehat{q}_1(\boldsymbol{f}) & \propto p(\boldsymbol{f}|\widetilde{\boldsymbol{\theta}}, \boldsymbol{g}; \mathcal{M}) \\ \widehat{q}_2(\boldsymbol{\theta}) & \propto p(\boldsymbol{\theta}|\boldsymbol{f}, \boldsymbol{g}; \mathcal{M}) \end{cases} \longrightarrow \begin{cases} \text{Prise en compte des incertitudes} \\ \text{de } \widehat{\boldsymbol{\theta}} \text{ pour } \widehat{\boldsymbol{f}} \text{ et vise versa.} \end{cases}$$

Hierarchical models and hidden variables

- All the mixture models and some of simple models can be modeled via hidden variables $\boldsymbol{z}$.

$$p(f)=\sum_{k=1}^{K}\alpha_k p_k(f) \longrightarrow \left\{ \begin{array}{l} p(f|z=k)=p_k(f), \\ P(z=k)=\alpha_k, \quad \sum_k \alpha_k=1 \end{array} \right.$$

- Example 1: MoG model: $p_k(f)=\mathcal{N}(f|m_k,v_k)$
  2 Gaussians: $p_0=\mathcal{N}(0,v_0), p_1=\mathcal{N}(0,v_1), \alpha_0=\lambda, \alpha_1=1-\lambda$

$$p(f_j|\lambda,v_1,v_0)=\lambda\mathcal{N}(f_j|0,v_1)+(1-\lambda)\mathcal{N}(f_j|0,v_0)$$

$$\left\{ \begin{array}{l} p(f_j|z_j=0,v_0)=\mathcal{N}(f_j|0,v_0), \\ p(f_j|z_j=1,v_1)=\mathcal{N}(f_j|0,v_1), \end{array} \right. \text{ and } \left\{ \begin{array}{l} P(z_j=0)=\lambda, \\ P(z_j=1)=1-\lambda \end{array} \right.$$

$$\left\{ \begin{array}{l} p(\boldsymbol{f}|\boldsymbol{z})=\prod_j p(f_j|z_j)=\prod_j \mathcal{N}\left(f_j|0,v_{z_j}\right) \propto \exp\left\{-\frac{1}{2}\sum_j \frac{f_j^2}{v_{z_j}}\right\} \\ p(\boldsymbol{z})=\lambda^{n_1}(1-\lambda)^{n_0}, \qquad n_0=\sum_j \delta(z_j), \quad n_1=\sum_j \delta(z_j-1) \end{array} \right.$$

Hierarchical models and hidden variables

- Example 2: Student-t model

$$\mathcal{S}t(f|\nu) \propto \exp\left\{-\frac{\nu+1}{2}\log\left(1+f^2/\nu\right)\right\}$$

- Infinite mixture

$$\mathcal{S}t(f|\nu) \propto= \int_0^\infty \mathcal{N}(f|,0,1/z)\,\mathcal{G}(z|\alpha,\beta)\,\mathsf{d}z, \quad \text{with } \alpha=\beta=\nu/2$$

$$\left\{\begin{array}{lll} p(\boldsymbol{f}|\boldsymbol{z}) & = \prod_j p(f_j|z_j) = \prod_j \mathcal{N}(f_j|0,1/z_j) \propto \exp\left\{-\frac{1}{2}\sum_j z_j f_j^2\right\} \\ p(\boldsymbol{z}|\alpha,\beta) & = \prod_j \mathcal{G}(z_j|\alpha,\beta) \propto \prod_j {z_j}^{(\alpha-1)} \exp\{-\beta z_j\} \\ & \propto \exp\left\{\sum_j (\alpha-1)\ln z_j - \beta z_j\right\} \\ p(\boldsymbol{f},\boldsymbol{z}|\alpha,\beta) & \propto \exp\left\{-\frac{1}{2}\sum_j z_j f_j^2 + (\alpha-1)\ln z_j - \beta z_j\right\} \end{array}\right.$$

Hierarchical models and hidden variables

- Example 3: Laplace (Double Exponential) model

$$\mathcal{DE}(f|a) = \frac{a}{2} \exp\{-a|f|\} = \int_0^\infty \mathcal{N}(f|,0,z)\,\mathcal{E}(z|a^2/2)\,\mathrm{d}z, \quad a > 0$$

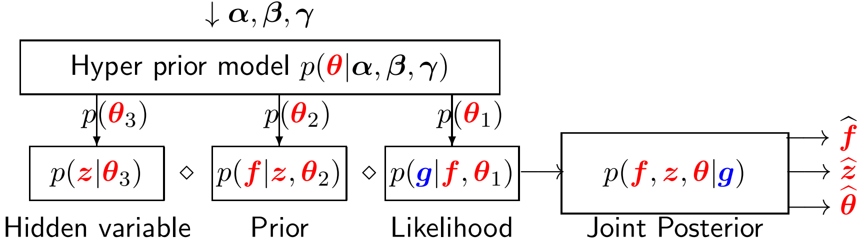
$$\begin{cases} p(\boldsymbol{f}|\boldsymbol{z}) & = \prod_j p(f_j|z_j) = \prod_j \mathcal{N}(f_j|0,z_j) \propto \exp\left\{-\frac{1}{2}\sum_j f_j^2/z_j\right\} \\ p(\boldsymbol{z}|\frac{a^2}{2}) & = \prod_j \mathcal{E}(z_j|\frac{a^2}{2}) \propto \exp\left\{\sum_j \frac{a^2}{2} z_j\right\} \\ p(\boldsymbol{f},\boldsymbol{z}|\frac{a^2}{2}) & \propto \exp\left\{-\frac{1}{2}\sum_j f_j^2/z_j + \frac{a^2}{2} z_j\right\} \end{cases}$$

- With these models we have:

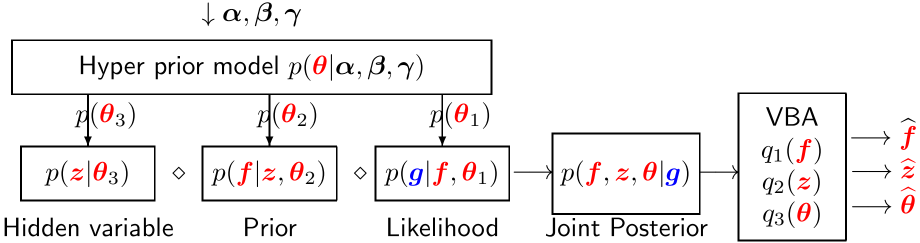
$$p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\,p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta}_2)\,p(\boldsymbol{z}|\boldsymbol{\theta}_3)\,p(\boldsymbol{\theta})$$

Summary of Bayesian estimation 3

- Full Bayesian Hierarchical Model with Hyperparameter Estimation



- Full Bayesian Hierarchical Model and Variational Approximation



Bayesian Computation and Algorithms

- Often, the expression of $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})$ is complex.
- Its optimization (for Joint MAP) or
its marginalization or integration (for Marginal MAP or PM)
is not easy
- Two main techniques:
MCMC and Variational Bayesian Approximation (VBA)
- MCMC:
Needs the expressions of the conditionals
$p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g})$, $p(\boldsymbol{z}|\boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{g})$, and $p(\boldsymbol{\theta}|\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{g})$
- VBA: Approximate $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})$ by a separable one

$$q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f})\, q_2(\boldsymbol{z})\, q_3(\boldsymbol{\theta})$$

and do any computations with these separable ones.

Bayesian Variational Approximation

- Objective: Approximate $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g})$ by a separable one $q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f})\, q_2(\boldsymbol{f})\, q_3(\boldsymbol{\theta})$
- Criterion:

$$\mathrm{KL}(q:p) = \int q \ln \frac{q}{p} = \left\langle \ln \frac{q}{p} \right\rangle_q$$

- Free energy: $\mathrm{KL}(q:p) = \ln p(\boldsymbol{g}|\mathcal{M}) - \mathcal{F}(q)$ where:

$$p(\boldsymbol{g}|\mathcal{M}) = \int\int\int p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g}|\mathcal{M})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{z}\, \mathrm{d}\boldsymbol{\theta}$$

  with $p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g}|\mathcal{M}) = p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta})\, p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta})\, p(\boldsymbol{z}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ and $\mathcal{F}(q)$ is the free energy associated to q defined as

$$\mathcal{F}(q) = \left\langle \ln \frac{p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g}|\mathcal{M})}{q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta})} \right\rangle_q$$

- For a given model $\mathcal{M}$, minimizing $\mathrm{KL}(q:p)$ is equivalent to maximizing $\mathcal{F}(q)$ and when optimized, $\mathcal{F}(q^*)$ gives a lower bound for $\ln p(\boldsymbol{g}|\mathcal{M})$.

BVA with Scale Mixture Model of Student-t priors

Scale Mixture Model of Student-t:

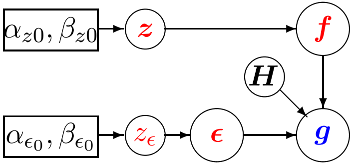
$$\mathcal{S}t(f_j|\nu) = \int_0^\infty \mathcal{N}(f_j|0, 1/z_j)\,\mathcal{G}(z_j|\nu/2, \nu/2)\,\mathrm{d}z_j$$

Hidden variables z_j:

$$\begin{aligned} p(\boldsymbol{f}|\boldsymbol{z}) &= \prod_j p(f_j|z_j) = \prod_j \mathcal{N}(f_j|0, 1/z_j) \propto \exp\left\{-\frac{1}{2}\sum_j z_j f_j^2\right\} \\ p(z_j|\alpha, \beta) &= \mathcal{G}(z_j|\alpha, \beta) \propto z_j^{(\alpha-1)} \exp\{-\beta z_j\} \text{ with } \alpha = \beta = \nu/2 \end{aligned}$$

Cauchy model is obtained when $\nu = 1$:

- Graphical model:

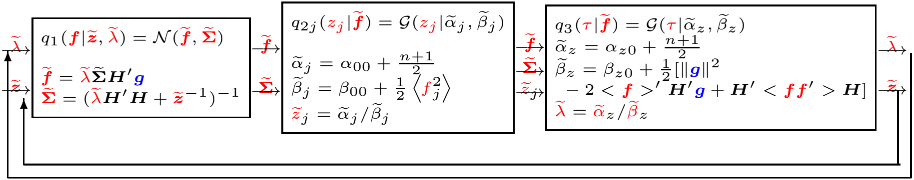


12. BVA with Student-t priors Algorithm

$$\begin{cases} p(\boldsymbol{g}|\boldsymbol{f}, z_\epsilon) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f}, (1/z_\epsilon)\boldsymbol{I}) \\ p(z_\epsilon|\alpha_{z0}, \beta_{z0}) = \mathcal{G}(z_\epsilon|\alpha_{z0}, \beta_{z0}) \\ p(\boldsymbol{f}|\boldsymbol{z}) = \prod_j \mathcal{N}\left(f_j|0, 1/z_j\right) \\ p(\boldsymbol{z}|\alpha_0, \beta_0) = \prod_j \mathcal{G}(z_j|\alpha_0, \beta_0) \\ \\ q_1(\boldsymbol{f}|\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}) = \mathcal{N}(\boldsymbol{f}|\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}) \\ \widetilde{\boldsymbol{\mu}} = \langle\lambda\rangle\, \widetilde{\boldsymbol{\Sigma}}\boldsymbol{H}'\boldsymbol{g} \\ \widetilde{\boldsymbol{\Sigma}} = (\langle\lambda\rangle\, \boldsymbol{H}'\boldsymbol{H} + \widetilde{\boldsymbol{Z}})^{-1}, \\ \text{with} \quad \widetilde{\boldsymbol{Z}} = \widetilde{\boldsymbol{z}}^{-1} = \text{diag}\,[\widetilde{\boldsymbol{z}}] \end{cases}$$

$$\begin{cases} q_{2j}(z_j) = \mathcal{G}(z_j|\widetilde{\alpha}_j, \widetilde{\beta}_j) \\ \widetilde{\alpha}_j = \alpha_{00} + 1/2 \\ \widetilde{\beta}_j = \beta_{00} + < f_j^2 > /2 \\ \\ q_3(z_\epsilon) = \mathcal{G}(z_\epsilon|\widetilde{\alpha}_{z\epsilon}, \widetilde{\beta}_{z\epsilon}), \\ \widetilde{\alpha}_{z\epsilon} = \alpha_{z0} + (n+1)/2 \\ \widetilde{\beta}_{z\epsilon} = \beta_{z0} + 1/2[\|\boldsymbol{g}\|^2 \\ -2\,\langle\boldsymbol{f}\rangle'\,\boldsymbol{H}'\boldsymbol{g} + \boldsymbol{H}'\,\langle\boldsymbol{f}\boldsymbol{f}'\rangle\,\boldsymbol{H}] \end{cases}$$

$$\begin{cases} < \boldsymbol{f} >= \widetilde{\boldsymbol{\mu}} \\ < \boldsymbol{f}\boldsymbol{f}' >= \widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}' \\ < f_j^2 >= [\widetilde{\boldsymbol{\Sigma}}]_{jj} + \widetilde{\mu}_j^2 \\ \widetilde{\lambda} = \widetilde{\alpha}_z/\widetilde{\beta}_z \\ \widetilde{z}_j = \widetilde{\alpha}_j/\widetilde{\beta}_j \end{cases}$$



13. Implementation issues

- In inverse problems, often we do not have access directly to the matrix $\boldsymbol{H}$. But, we can compute:
  - Forward operator : $\boldsymbol{H} \boldsymbol{f} \longrightarrow \boldsymbol{g}$ `g=direct(f,...)`
  - Adjoint operator : $\boldsymbol{H}' \boldsymbol{g} \longrightarrow \boldsymbol{f}$ `f=transp(g,...)`
- For any particular application, we can always write two programs (`direct` & `transp`) corresponding to the application of these two operators.
- To compute $\widetilde{\boldsymbol{f}}$, we use a gradient based optimization algorithm which will use these operators.
- We may also need to compute the diagonal elements of $[\boldsymbol{H}'\boldsymbol{H}]$.. We also developped algorithms which computes these diagonal elements with the same programs (`direct` & `transp`)

14. Conclusions and Perspectives

- We proposed a list of different probabilistic prior models which can be used for sparsity enforcing.
- We classified these models in two categories: simple heavy tails and hierarchical mixture models
- We showed how to use these models for inverse problems where the desired solutions are sparse
- Different algorithms have been developed and their relative performances are compared.
- We use these models for inverse problems in different signal and image processing applications such as:
  - Period estimation in biological time series
  - X ray Computed Tomography,
  - Signal deconvolution in Proteomic and molecular imaging
  - Diffraction Optical Tomography
  - Microwave Imaging, Acoustic imaging and sources localization
  - Synthetic Aperture Radar (SAR) Imaging

15. References

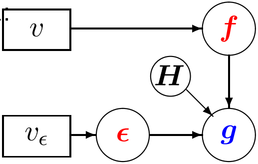
1. A. Mohammad-Djafari, "Bayesian approach with prior models which enforce sparsity in signal and image processing," EURASIP Journal on Advances in Signal Processing, vol. Special issue on Sparse Signal Processing, (2012).
2. S. Zhu, A. Mohammad-Djafari, H. Wang, B. Deng, X. Li and J. Mao J, "Parameter estimation for SAR micromotion target based on sparse signal representation," EURASIP Journal on Advances in Signal Processing, vol. Special issue on Sparse Signal Processing, (2012).
3. N. Chu, J. Picheral and A. Mohammad-Djafari, "A robust super-resolution approach with sparsity constraint for near-field wideband acoustic imaging," *IEEE International Symposium on Signal Processing and Information Technology* pp 286–289, Bilbao, Spain, Dec14-17,2011
4. N. Bali and A. Mohammad-Djafari, "Bayesian Approach With Hidden Markov Modeling and Mean Field Approximation for Hyperspectral Data Analysis," IEEE Trans. on Image Processing 17: 2. 217-225 Feb. (2008).
5. J. Griffin and P. Brown, "Inference with normal-gamma prior distributions in regression problems," *Bayesian Analysis*, 2010.
6. N. Polson and J. Scott., "Shrink globally, act locally: sparse Bayesian regularization and prediction," *Bayesian Statistics 9*, 2010.
7. T. Park and G. Casella., "The Bayesian Lasso," *Journal of the American Statistical Association*, 2008.
8. C. Févotte and S. Godsill, "A Bayesian aproach for blind separation of sparse source," *IEEE Transactions on Audio, Speech, and Language processing*, 2006.
9. H. Snoussi and J. Idier., "Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures," *IEEE Trans. on Signal Processing*, 2006.
10. J. R. H. Ishwaran, "Spike and slab variable selection: Frequentist and Bayesian strategies," *Annals of Statistics*, 2005.
11. M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, 2001.

14. References ..

1. Abib. Doucet and P. Duvaut, "Bayesian estimation of state-space models applied to deconvolution of Bernoulli-Gaussian processes," *Signal Processing*, vol. 57, no. 2, 1997.
2. P. Williams, "Bayesian regularization and pruning using a Laplace prior," *Neural Computation*, 1995.
3. M. Lavielle, "Bayesian deconvolution of Bernoulli-Gaussian processes," *Signal Processing*, vol. 33, pp. 67–79, 1993.
4. T. Mitchell and J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 1988.
5. J. J. Kormylo and J. M. Mendel, "Maximum-likelihood detection and estimation of Bernoulli-Gaussian processes," vol. 28, pp. 482–488, 1982.
6. H. Snoussi and A. Mohammad-Djafari, " Estimation of Structured Gaussian Mixtures: The Inverse EM Algorithm," IEEE Trans. on Signal Processing 55: 7. 3185-3191 July (2007).
7. N. Bali and A. Mohammad-Djafari, "A variational Bayesian Algorithm for BSS Problem with Hidden Gauss-Markov Models for the Sources," in: Independent Component Analysis and Signal Separation (ICA 2007) Edited by:M.E. Davies, Ch.J. James, S.A. Abdallah, M.D. Plumbley. 137-144 Springer (LNCS 4666) (2007).
8. N. Bali and A. Mohammad-Djafari, "Hierarchical Markovian Models for Joint Classification, Segmentation and Data Reduction of Hyperspectral Images" ESANN 2006, September 4-8, Belgium. (2006)
9. M. Ichir and A. Mohammad-Djafari, "Hidden Markov models for wavelet-based blind source separation," IEEE Trans. on Image Processing 15: 7. 1887-1899 July (2005)
10. S. Moussaoui, C. Carteret, D. Brie and A Mohammad-Djafari, "Bayesian analysis of spectral mixture data using Markov Chain Monte Carlo methods sampling," Chemometrics and Intelligent Laboratory Systems 81: 2. 137-148 (2005).
11. H. Snoussi and A. Mohammad-Djafari, "Fast joint separation and segmentation of mixed images" Journal of Electronic Imaging 13: 2. 349-361 April (2004)
12. H. Snoussi and A. Mohammad-Djafari, "Bayesian unsupervised learning for source separation with mixture of Gaussians prior," Journal of VLSI Signal Processing Systems 37: 2/3. 263-279 June/July (2004)

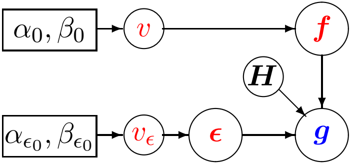
Separable models

- Model 1:



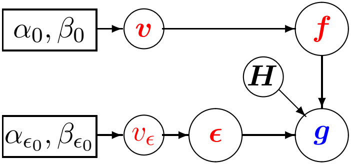
$$p(f_j|v) = \mathcal{N}(0, v)$$
$$p(\boldsymbol{f}|v) \propto \exp\left\{-\frac{1}{2}\sum_j \frac{f_j^2}{v}\right\}$$
$$p(\boldsymbol{g}|\boldsymbol{f}, v_\epsilon) = \mathcal{N}(\boldsymbol{H}\boldsymbol{f}, v_\epsilon \boldsymbol{I})$$

- Model 2:



$$p(f_j|v) = \mathcal{N}(0, v)$$
$$p(\boldsymbol{f}|v) \propto \exp\left\{-\frac{1}{2}\sum_j \frac{f_j^2}{v}\right\}$$
$$p(\boldsymbol{g}|\boldsymbol{f}, v_\epsilon) = \mathcal{N}(\boldsymbol{H}\boldsymbol{f}, v_\epsilon \boldsymbol{I})$$
$$p(v_\epsilon|\alpha_{\epsilon 0}, \beta_{\epsilon 0}) = \mathcal{G}(\alpha_{\epsilon 0}, \beta_{\epsilon 0})$$
$$p(v|\alpha_0, \beta_0) = \mathcal{G}(\alpha_0, \beta_0)$$

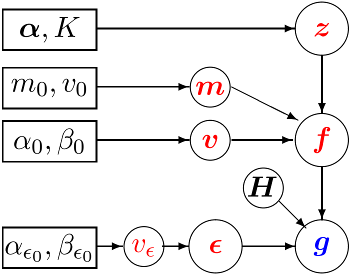
- Model 3:



$$p(f_j|v_j) = \mathcal{N}(0, v_j)$$
$$p(\boldsymbol{f}|\boldsymbol{v}) \propto \exp\left\{-\frac{1}{2}\sum_j \frac{f_j^2}{v_j}\right\}$$
$$p(v_j|\alpha_0, \beta_0) = \mathcal{G}(\alpha_0, \beta_0)$$

Separable models

- Model 4:



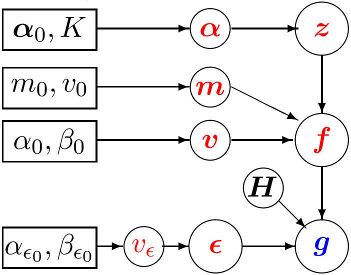
$$p(z_j = k|\alpha_k) = \alpha_k, \quad \sum_k \alpha_k = 1$$

$$p(f_j|z_j = k) = \mathcal{N}(m_{j\,k}, v_{j\,k})$$
$$p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{v}) = \sum_j \mathcal{N}(m_{z\,j}, v_{z\,j})$$

$$p(\boldsymbol{g}|\boldsymbol{f}, v_\epsilon) = \mathcal{N}(\boldsymbol{H}\boldsymbol{f}, v_\epsilon \boldsymbol{I})$$
$$p(v_\epsilon|\alpha_{\epsilon 0}, \beta_{\epsilon 0}) = \mathcal{IG}(\alpha_{\epsilon 0}, \beta_{\epsilon 0})$$

Separale models

- Model 5:



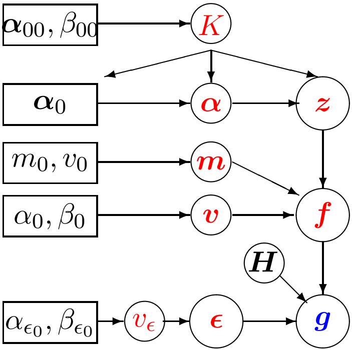
$$p(z_j = k|\alpha_k) = \alpha_k, \quad \sum_k \alpha_k = 1$$

$$p(f_j|z_j = k) = \mathcal{N}(m_{j_k}, v_{j_k})$$
$$p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{v}) = \sum_j \mathcal{N}(m_{zj}, v_{zj})$$

$$p(\boldsymbol{g}|\boldsymbol{f}, v_\epsilon) = \mathcal{N}(\boldsymbol{H}\boldsymbol{f}, v_\epsilon \boldsymbol{I})$$
$$p(v_\epsilon|\alpha_{\epsilon 0}, \beta_{\epsilon 0}) = \mathcal{IG}(\alpha_{\epsilon 0}, \beta_{\epsilon 0})$$

Separale models

- Model 6:



$$p(z_j = k|\alpha_k) = \alpha_k, \quad \sum_k \alpha_k = 1$$

$$p(f_j|z_j = k) = \mathcal{N}(m_{j_k}, v_{j_k})$$
$$p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{v}) = \sum_j \mathcal{N}(m_{zj}, v_{zj})$$

$$p(\boldsymbol{g}|\boldsymbol{f}, v_\epsilon) = \mathcal{N}(\boldsymbol{H}\boldsymbol{f}, v_\epsilon \boldsymbol{I})$$
$$p(v_\epsilon|\alpha_{\epsilon 0}, \beta_{\epsilon 0}) = \mathcal{IG}(\alpha_{\epsilon 0}, \beta_{\epsilon 0})$$